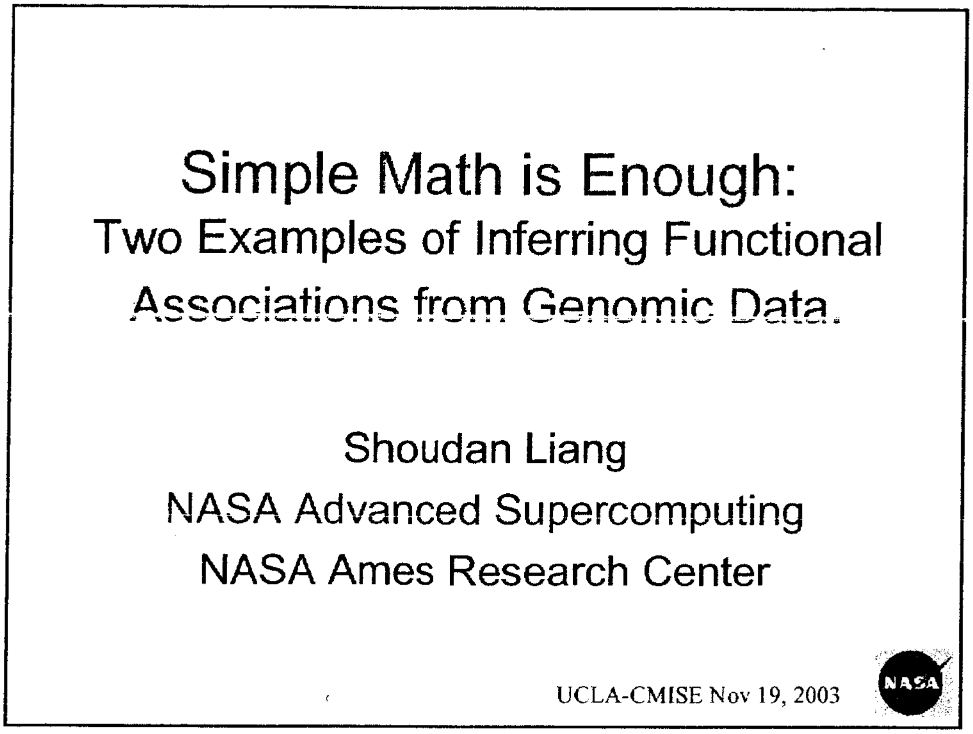
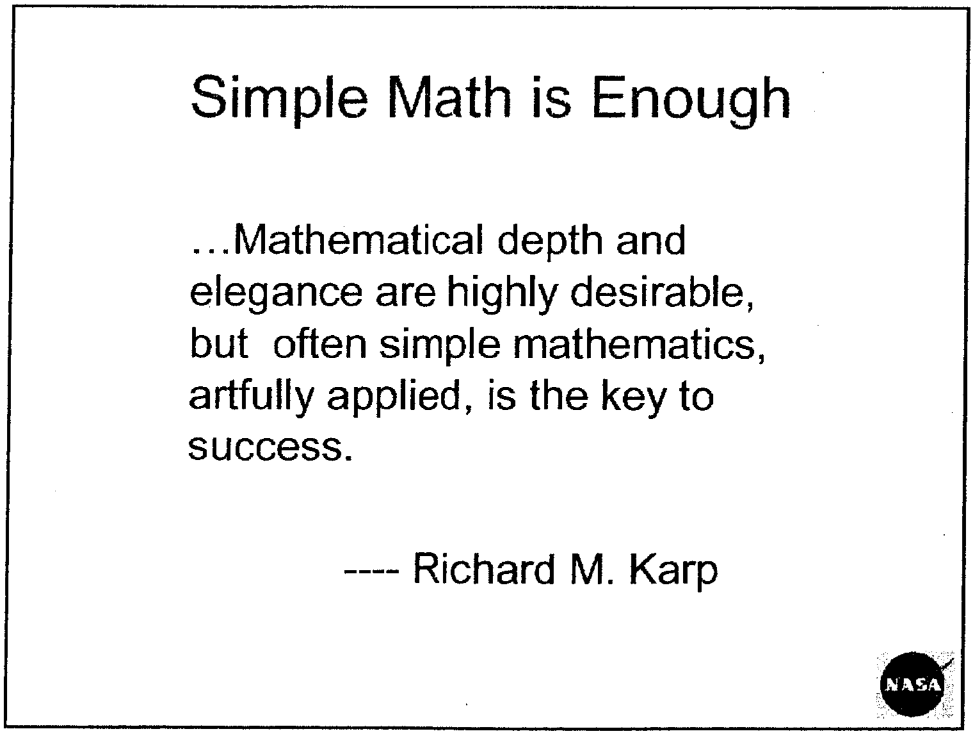


Simple Math is Enough:

Two Examples of Inferring Functional Associations from Genomic Data.

Shoudan Liang
NASA Advanced Supercomputing
NASA Ames Research Center

UCLA-CMISE Nov 19, 2003



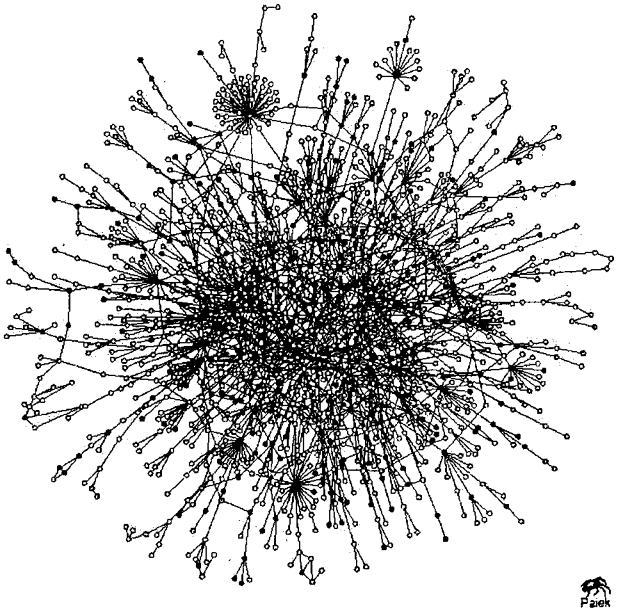
Simple Math is Enough

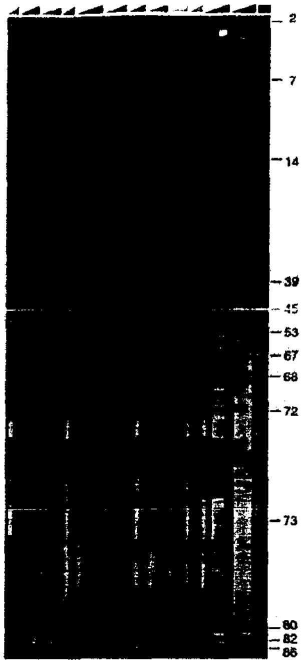
...Mathematical depth and elegance are highly desirable, but often simple mathematics, artfully applied, is the key to success.

---- Richard M. Karp

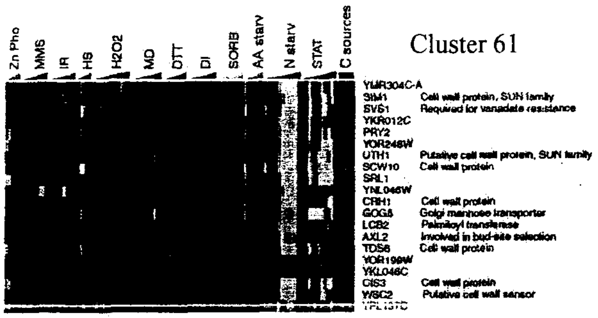
meaning making of genomic data

- Genomic data
  - Two-hybrid protein-protein interactions
  - DNA microarray mRNA transcription
- High rate of error in current technologies
- Think some aspect of data that are both non-random and biologically meaningful
- Compute a p-value associated with such non-random feature and use it to weed out the false positive errors





Cluster 61



Clustering: a group of genes is selected based on the similarity in their expression under different stimulations

Common motifs: search for shared nucleotide motifs in DNA sequences a few hundred bp before the transcription start of each gene

Example: CRCGAAA R=A or G

Gasch AP and Eisen MB
Genome Biology 3(11), 1-22.

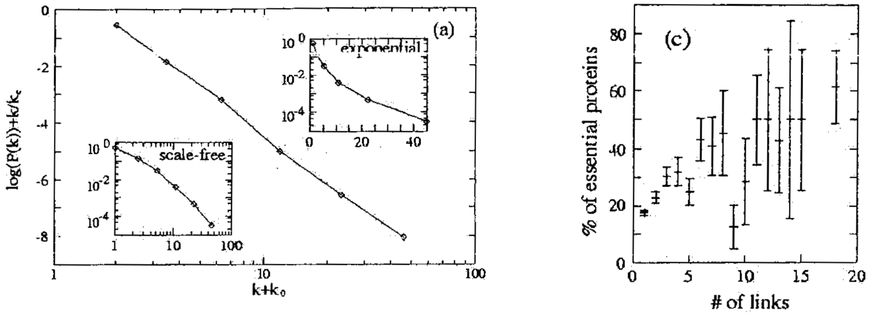
Tavazoie *et al. Nat. Gen.* 22 213.

meaning making of genomic data

- Genomic data
  - Two-hybrid protein-protein interactions
  - DNA microarray mRNA transcription
- High rate of error in current technologies
- Think some aspect of data that are both non-random and biologically meaningful
- Compute a p-value associated with such non-random feature and use it to weed out the false positive errors

Protein-protein interactions: non-random features

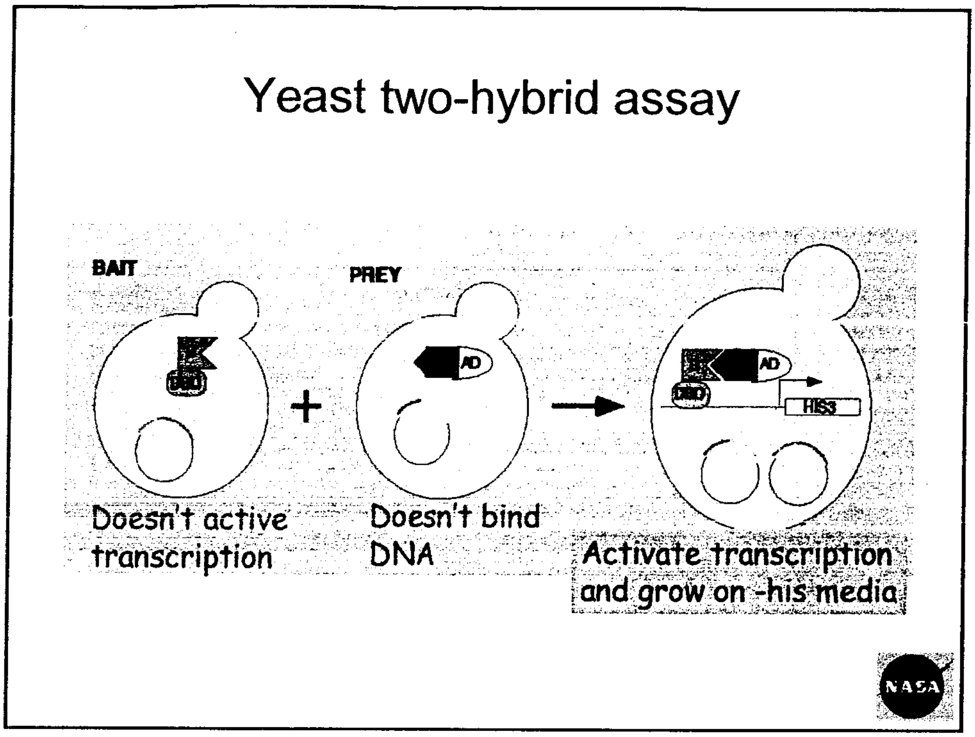
$$N_k = k^{\alpha}$$



Jeong et al., Nature (2001) 411:41-2.

In this talk…

- A method of suggesting protein functions based on protein-protein interaction data.
    - Samanta, M., Liang, S, *Proc Natl Acad Sci USA.* (2003) **100**, 12579-12583.
- A method of extracting protein-binding DNA motifs from a single microarray experiment.
    - Bussemaker *et al. Nat. Genet.* (2001) **27** 167-171.
    - Work in progress

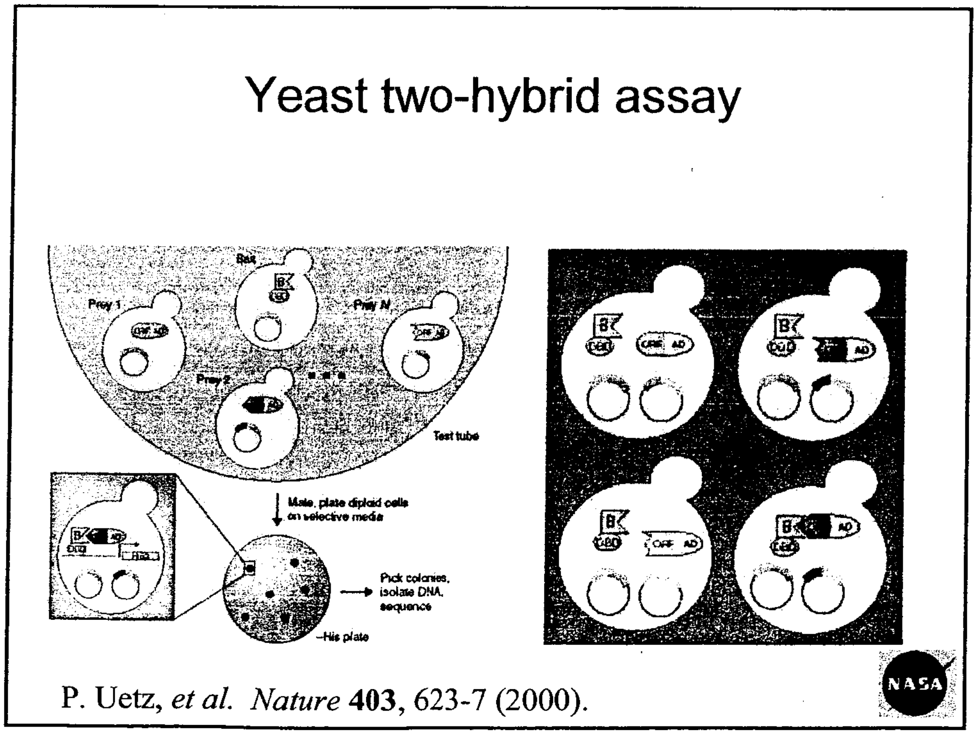


Yeast two-hybrid assay

Doesn't active transcription

Doesn't bind DNA

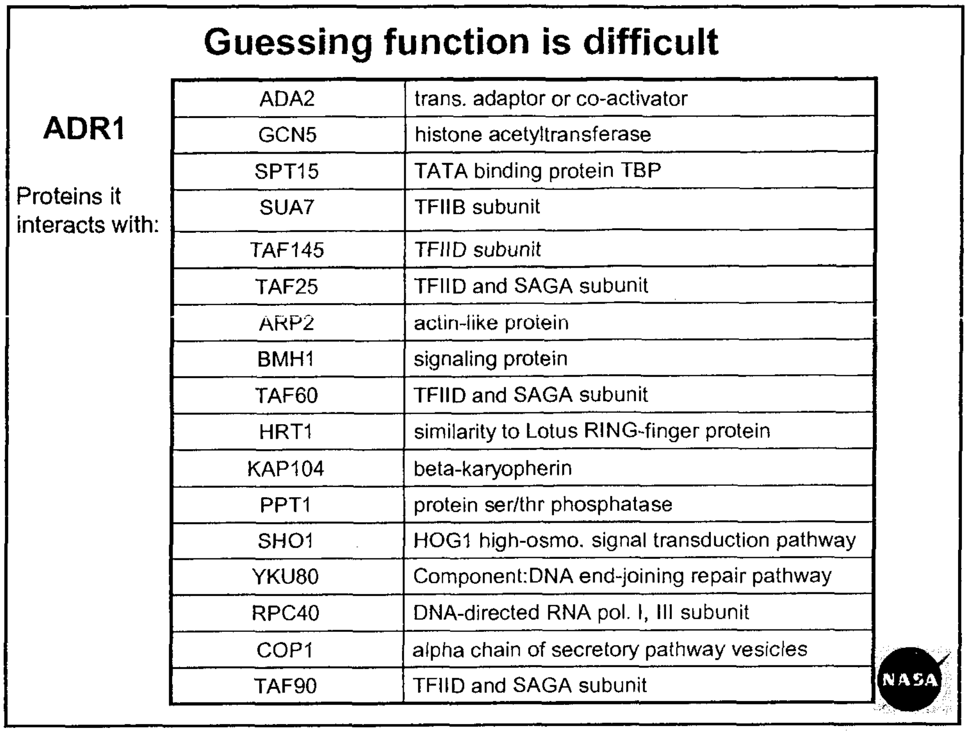
Activate transcription and grow on -his media



Yeast two-hybrid assay

Bait
Prey 1
Prey 2
Prey N
B
DBD
ORF AD
AD
Test tube
Mate, plate diploid cells on selective media
Pick colonies, isolate DNA, sequence
–His plate
NASA

P. Uetz, *et al.* *Nature* **403**, 623-7 (2000).

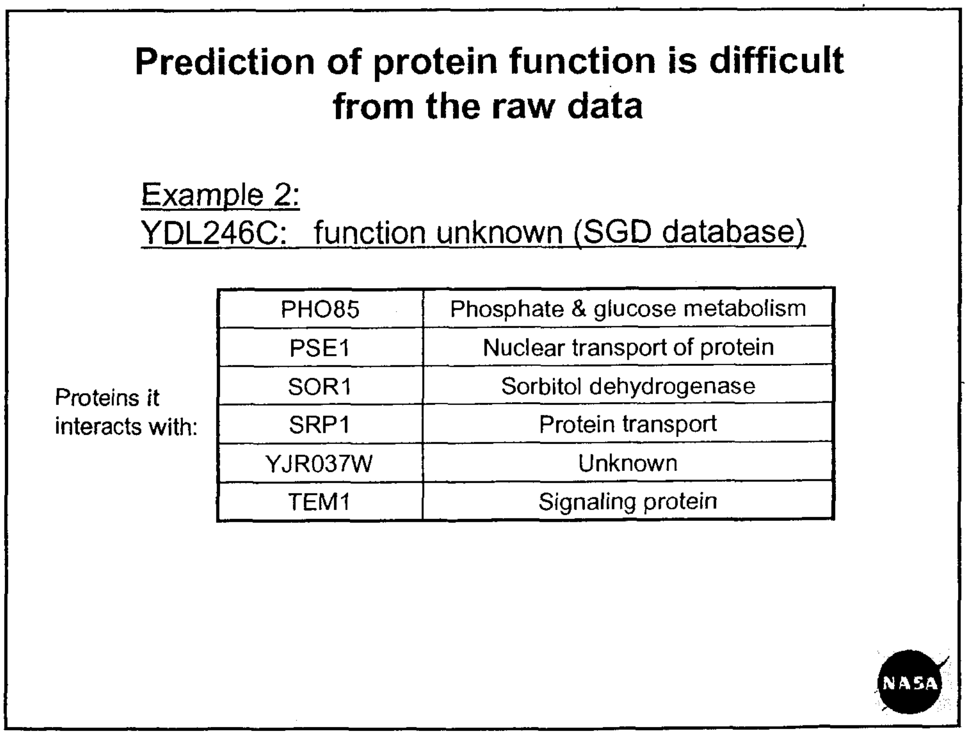


Guessing function is difficult

ADR1

Proteins it interacts with:

ADA2	trans. adaptor or co-activator
GCN5	histone acetyltransferase
SPT15	TATA binding protein TBP
SUA7	TFIIB subunit
TAF145	TFIID subunit
TAF25	TFIID and SAGA subunit
ARP2	actin-like protein
BMH1	signaling protein
TAF60	TFIID and SAGA subunit
HRT1	similarity to Lotus RING-finger protein
KAP104	beta-karyopherin
PPT1	protein ser/thr phosphatase
SHO1	HOG1 high-osmo. signal transduction pathway
YKU80	Component:DNA end-joining repair pathway
RPC40	DNA-directed RNA pol. I, III subunit
COP1	alpha chain of secretory pathway vesicles
TAF90	TFIID and SAGA subunit



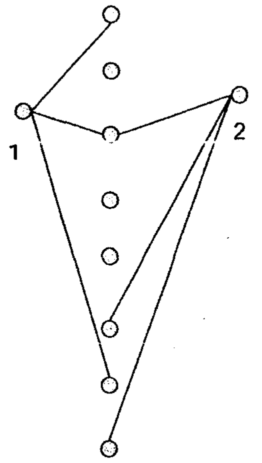
Prediction of protein function is difficult from the raw data

Example 2:
YDL246C: function unknown (SGD database)

Proteins it interacts with:

PHO85	Phosphate & glucose metabolism
PSE1	Nuclear transport of protein
SOR1	Sorbitol dehydrogenase
SRP1	Protein transport
YJR037W	Unknown
TEM1	Signaling protein

We derive p-value based on two proteins having a large number of interaction partners in common



Protein 1 interacts with n_1 partners; Protein 2 interacts with n_2 partners.

The probability *P* of having m partners in common

$$P = \frac{\binom{N}{m}\binom{N-m}{n_1-m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_1}\binom{N}{n_2}}$$

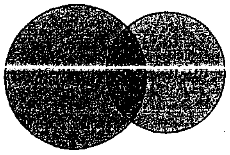
counting problem #1:

Distinct ways for protein 1 to have n_1 interacting partners is $\binom{N}{n_1} = \frac{N!}{(N-n_1)!n_1!}$

Similarly for protein 2 $\binom{N}{n_2} = \frac{N!}{(N-n_2)!n_2!}$

Total number of ways of having n_1 interacting partners for protein 1 *and* n_2 interacting partners for protein 2 $\binom{N}{n_1}\binom{N}{n_2}$

counting problem #2: The protein 1 and protein 2 have *m* interacting partners in common.

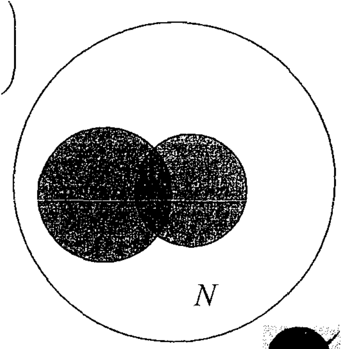


$$\binom{N}{m}\binom{N-m}{n_1-m}\binom{N-n_1}{n_2-m}$$

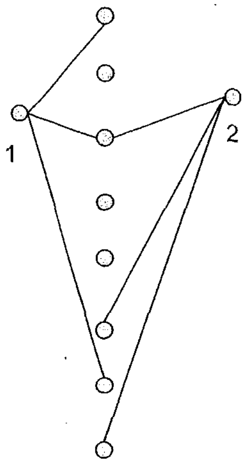
m common partners

n_1-*m* remaining partner for protein 1

n_2-*m* remaining partner for protein 2



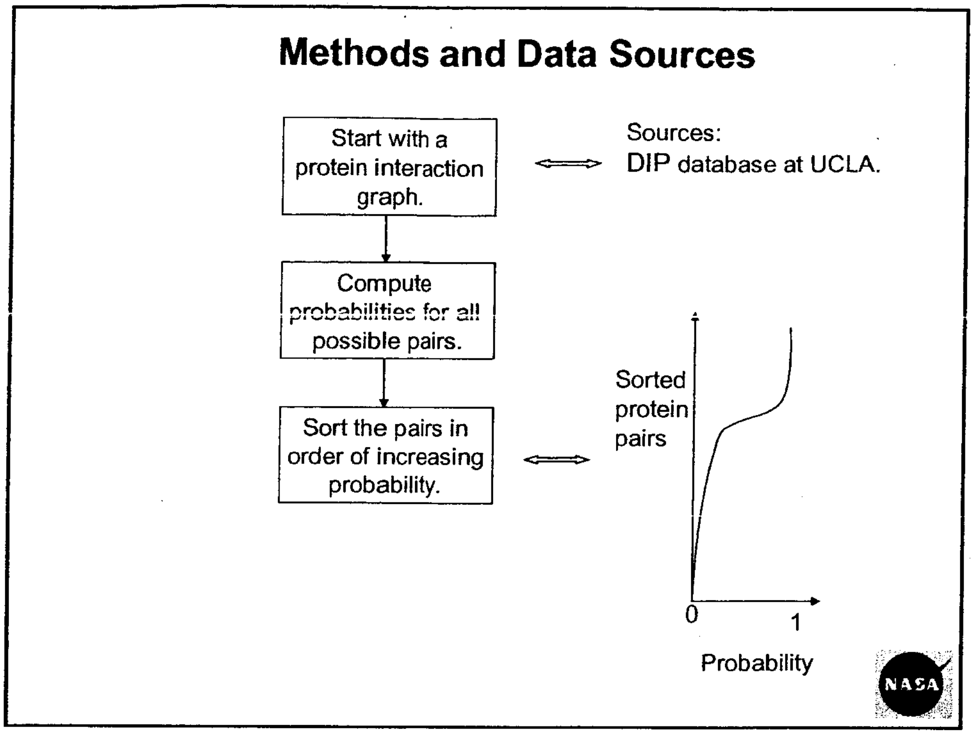
We derive p-value based on two proteins having a large number of interaction partners in common



Protein 1 interacts with n_1 partners; Protein 2 interacts with n_2 partners.

The probability *P* of having *m* partners in common

$$P=\frac{\binom{N}{m}\binom{N-m}{n_1-m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_1}\binom{N}{n_2}}$$



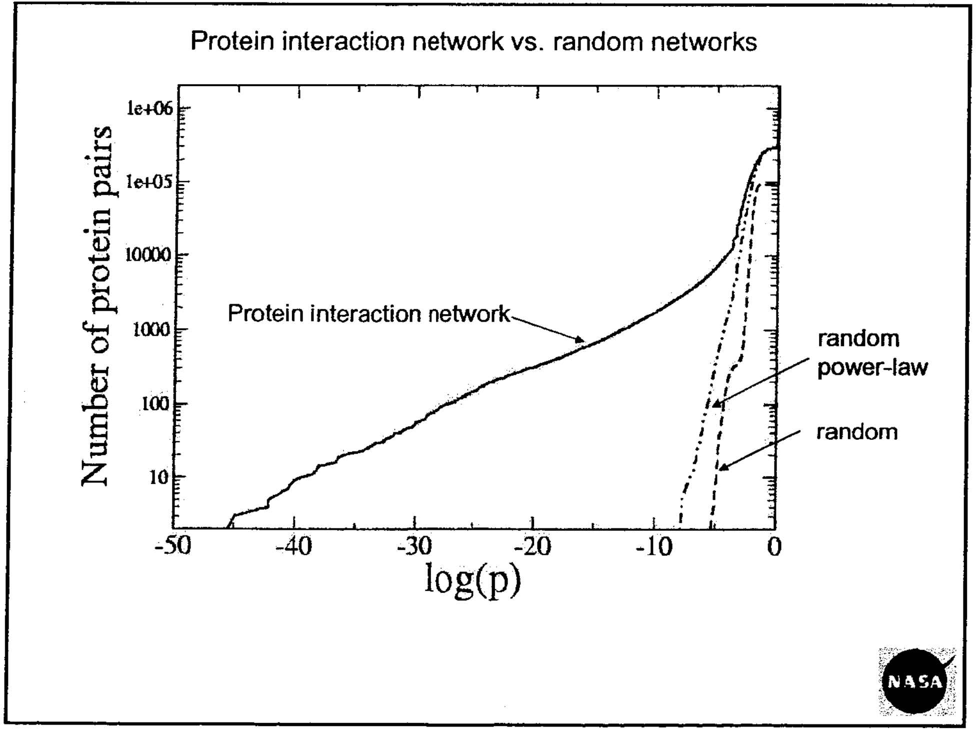
Methods and Data Sources

Start with a protein interaction graph.

Compute probabilities for all possible pairs.

Sort the pairs in order of increasing probability.

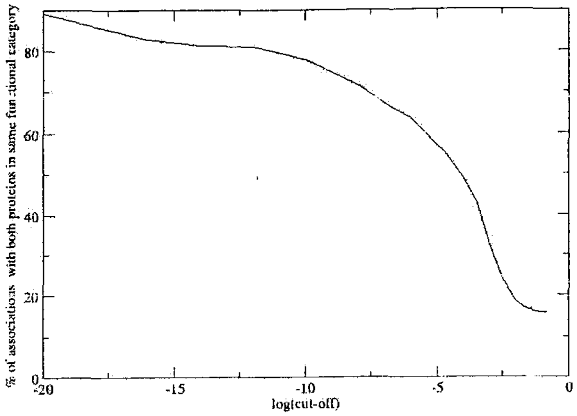
Sources:
DIP database at UCLA.



Protein interaction network vs. random networks

Top 1000 pairs are more than 70% likely to be have similar function of both proteins (random pair 3-6%).

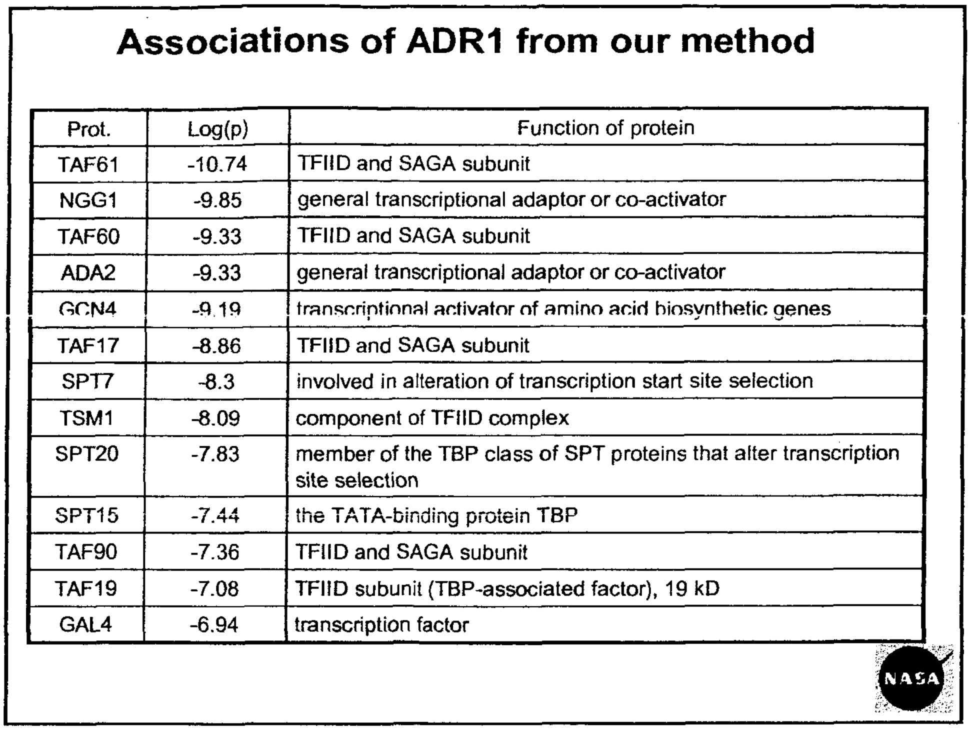
Functional similarities of protein pairs at different cut-offs



ADR1

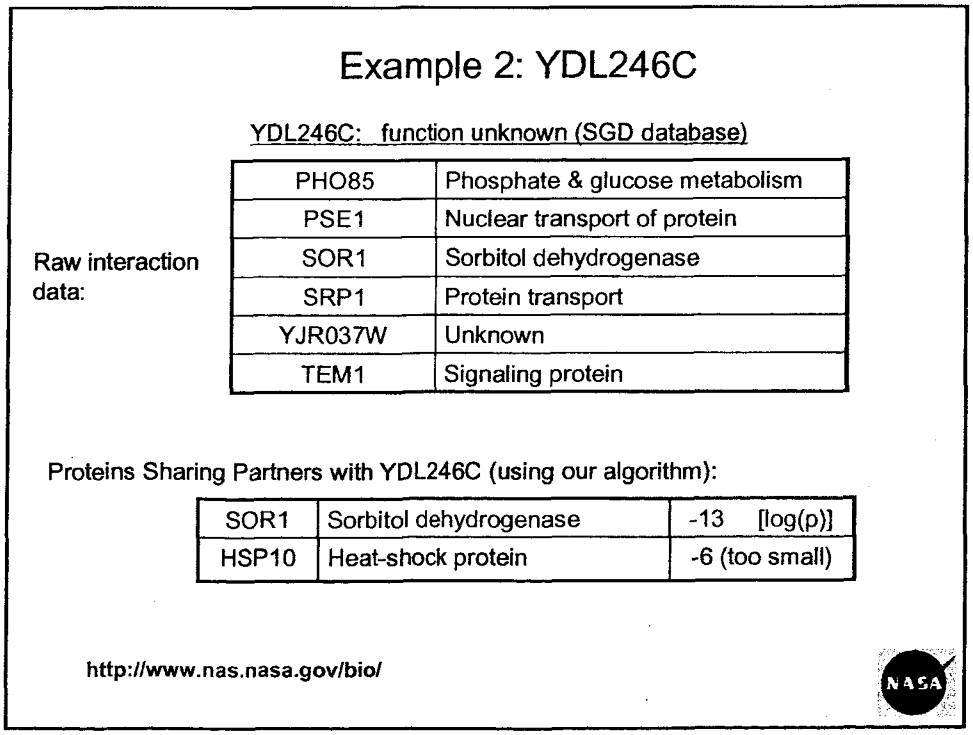
Raw interaction data (shown previously):

ADA2	trans. adaptor or co-activator
GCN5	histone acetyltransferase
SPT15	TATA binding protein TBP
SUA7	TFIIB subunit
TAF145	TFIID subunit
TAF25	TFIID and SAGA subunit
ARP2	actin-like protein
BMH1	signaling protein
TAF60	TFIID and SAGA subunit
HRT1	similarity to Lotus RING-finger protein
KAP104	beta-karyopherin
PPT1	protein ser/thr phosphatase
SHO1	HOG1 high-osmo. signal transduction pathway
YKU80	Component:DNA end-joining repair pathway
RPC40	DNA-directed RNA pol. I, III subunit
COP1	alpha chain of secretory pathway vesicles
TAF90	TFIID and SAGA subunit



Associations of ADR1 from our method

Prot.	Log(p)	Function of protein
TAF61	-10.74	TFIID and SAGA subunit
NGG1	-9.85	general transcriptional adaptor or co-activator
TAF60	-9.33	TFIID and SAGA subunit
ADA2	-9.33	general transcriptional adaptor or co-activator
GCN4	-9.19	transcriptional activator of amino acid biosynthetic genes
TAF17	-8.86	TFIID and SAGA subunit
SPT7	-8.3	involved in alteration of transcription start site selection
TSM1	-8.09	component of TFIID complex
SPT20	-7.83	member of the TBP class of SPT proteins that alter transcription site selection
SPT15	-7.44	the TATA-binding protein TBP
TAF90	-7.36	TFIID and SAGA subunit
TAF19	-7.08	TFIID subunit (TBP-associated factor), 19 kD
GAL4	-6.94	transcription factor



Example 2: YDL246C

YDL246C: function unknown (SGD database)

Raw interaction data:

PHO85	Phosphate & glucose metabolism
PSE1	Nuclear transport of protein
SOR1	Sorbitol dehydrogenase
SRP1	Protein transport
YJR037W	Unknown
TEM1	Signaling protein

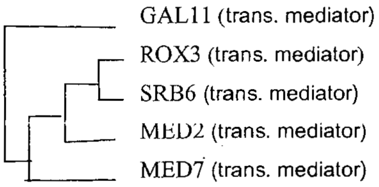
Proteins Sharing Partners with YDL246C (using our algorithm):

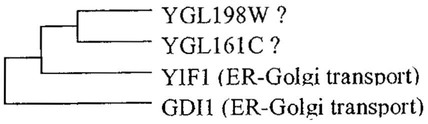
SOR1	Sorbitol dehydrogenase	-13 [log(p)]
HSP10	Heat-shock protein	-6 (too small)

http://www.nas.nasa.gov/bio/

By clustering we can recover complexes and pathways

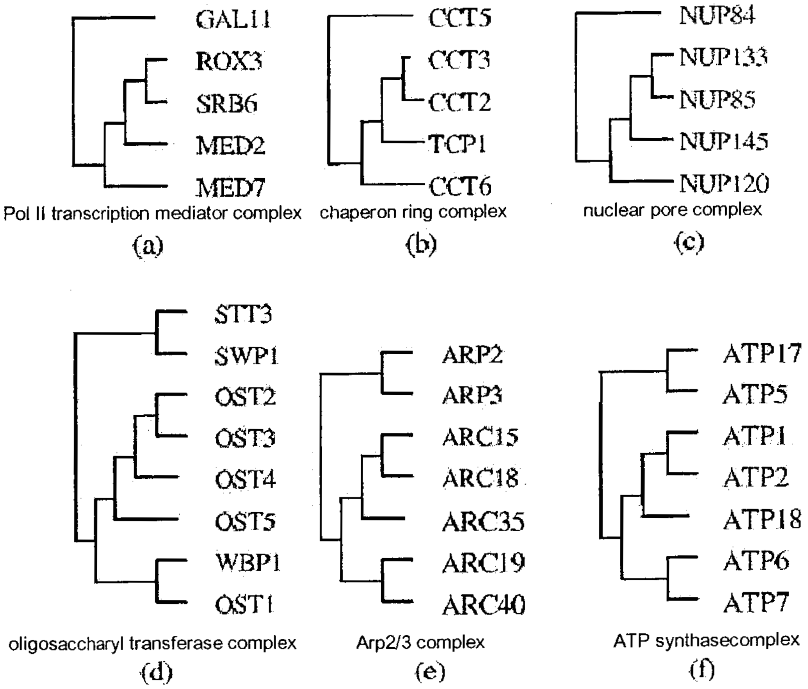
202 modules are reconstructed covering most aspects of cell.





We predicted functions of 81 unannotated proteins.
22 out 23 are now known to be correct.

YDL246C: same function as SOR1 (sorbitol dehydrogenease)



Pol II transcription mediator complex
(a)

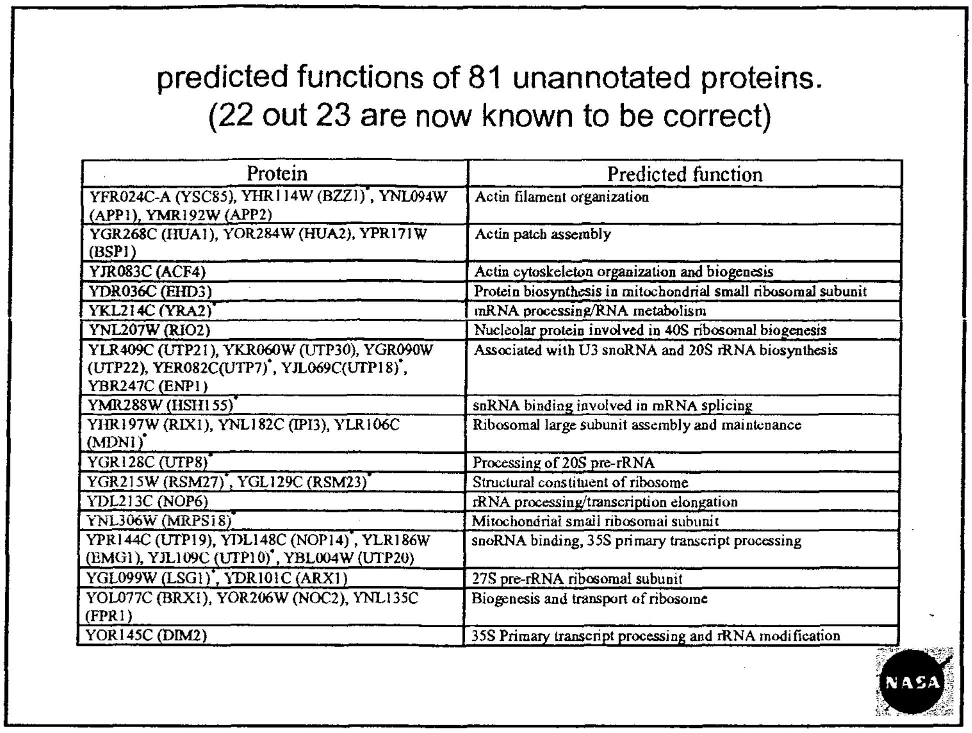
chaperon ring complex
(b)

nuclear pore complex
(c)

oligosaccharyl transferase complex
(d)

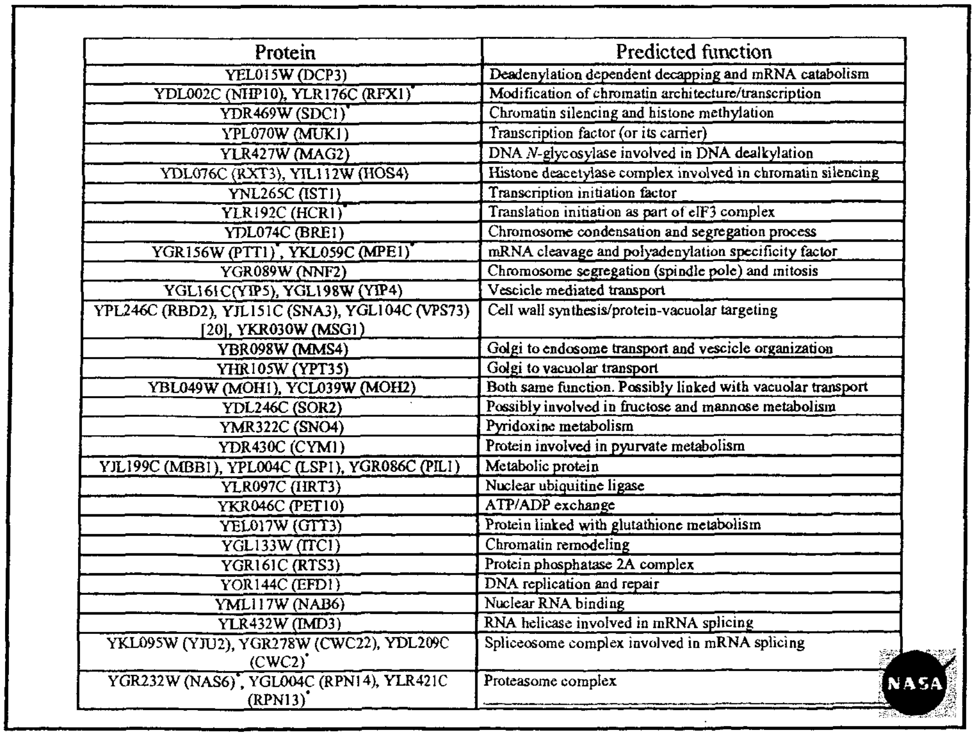
Arp2/3 complex
(e)

ATP synthasecomplex
(f)



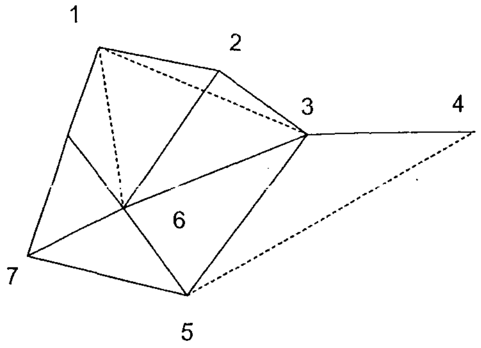
predicted functions of 81 unannotated proteins. (22 out 23 are now known to be correct)

Protein	Predicted function
YFR024C-A (YSC85), YHR114W (BZZ1)*, YNL094W (APP1), YMR192W (APP2)	Actin filament organization
YGR268C (HUA1), YOR284W (HUA2), YPR171W (BSP1)	Actin patch assembly
YJR083C (ACF4)	Actin cytoskeleton organization and biogenesis
YDR036C (EHD3)	Protein biosynthesis in mitochondrial small ribosomal subunit
YKL214C (YRA2)*	mRNA processing/RNA metabolism
YNL207W (RIO2)	Nucleolar protein involved in 40S ribosomal biogenesis
YLR409C (UTP21), YKR060W (UTP30), YGR090W (UTP22), YER082C(UTP7)*, YJL069C(UTP18)*, YBR247C (ENP1)	Associated with U3 snoRNA and 20S rRNA biosynthesis
YMR288W (HSH155)*	snRNA binding involved in mRNA splicing
YHR197W (RIX1), YNL182C (IPI3), YLR106C (MDN1)*	Ribosomal large subunit assembly and maintenance
YGR128C (UTP8)*	Processing of 20S pre-rRNA
YGR215W (RSM27)*, YGL129C (RSM23)*	Structural constituent of ribosome
YDL213C (NOP6)	rRNA processing/transcription elongation
YNL306W (MRPS18)*	Mitochondrial small ribosomal subunit
YPR144C (UTP19), YDL148C (NOP14)*, YLR186W (EMG1), YJL109C (UTP10)*, YBL004W (UTP20)	snoRNA binding, 35S primary transcript processing
YGL099W (LSG1)*, YDR101C (ARX1)	27S pre-rRNA ribosomal subunit
YOL077C (BRX1), YOR206W (NOC2), YNL135C (FPR1)	Biogenesis and transport of ribosome
YOR145C (DIM2)	35S Primary transcript processing and rRNA modification



Protein	Predicted function
YEL015W (DCP3)	Deadenylation dependent decapping and mRNA catabolism
YDL002C (NHP10), YLR176C (RFX1)*	Modification of chromatin architecture/transcription
YDR469W (SDC1)*	Chromatin silencing and histone methylation
YPL070W (MUK1)	Transcription factor (or its carrier)
YLR427W (MAG2)	DNA *N*-glycosylase involved in DNA dealkylation
YDL076C (RXT3), YIL112W (HOS4)	Histone deacetylase complex involved in chromatin silencing
YNL265C (IST1)	Transcription initiation factor
YLR192C (HCR1)*	Translation initiation as part of eIF3 complex
YDL074C (BRE1)	Chromosome condensation and segregation process
YGR156W (PTT1)*, YKL059C (MPE1)*	mRNA cleavage and polyadenylation specificity factor
YGR089W (NNF2)	Chromosome segregation (spindle pole) and mitosis
YGL161C(YIP5), YGL198W (YIP4)	Vescicle mediated transport
YPL246C (RBD2), YJL151C (SNA3), YGL104C (VPS73) [20], YKR030W (MSG1)	Cell wall synthesis/protein-vacuolar targeting
YBR098W (MMS4)	Golgi to endosome transport and vescicle organization
YHR105W (YPT35)	Golgi to vacuolar transport
YBL049W (MOH1), YCL039W (MOH2)	Both same function. Possibly linked with vacuolar transport
YDL246C (SOR2)	Possibly involved in fructose and mannose metabolism
YMR322C (SNO4)	Pyridoxine metabolism
YDR430C (CYM1)	Protein involved in pyurvate metabolism
YJL199C (MBB1), YPL004C (LSP1), YGR086C (PIL1)	Metabolic protein
YLR097C (HRT3)	Nuclear ubiquitine ligase
YKR046C (PET10)	ATP/ADP exchange
YEL017W (GTT3)	Protein linked with glutathione metabolism
YGL133W (ITC1)	Chromatin remodeling
YGR161C (RTS3)	Protein phosphatase 2A complex
YOR144C (EFD1)	DNA replication and repair
YML117W (NAB6)	Nuclear RNA binding
YLR432W (IMD3)	RNA helicase involved in mRNA splicing
YKL095W (YJU2), YGR278W (CWC22), YDL209C (CWC2)*	Spliceosome complex involved in mRNA splicing
YGR232W (NAS6)*, YGL004C (RPN14), YLR421C (RPN13)*	Proteasome complex

Our method is very robust from noise !!



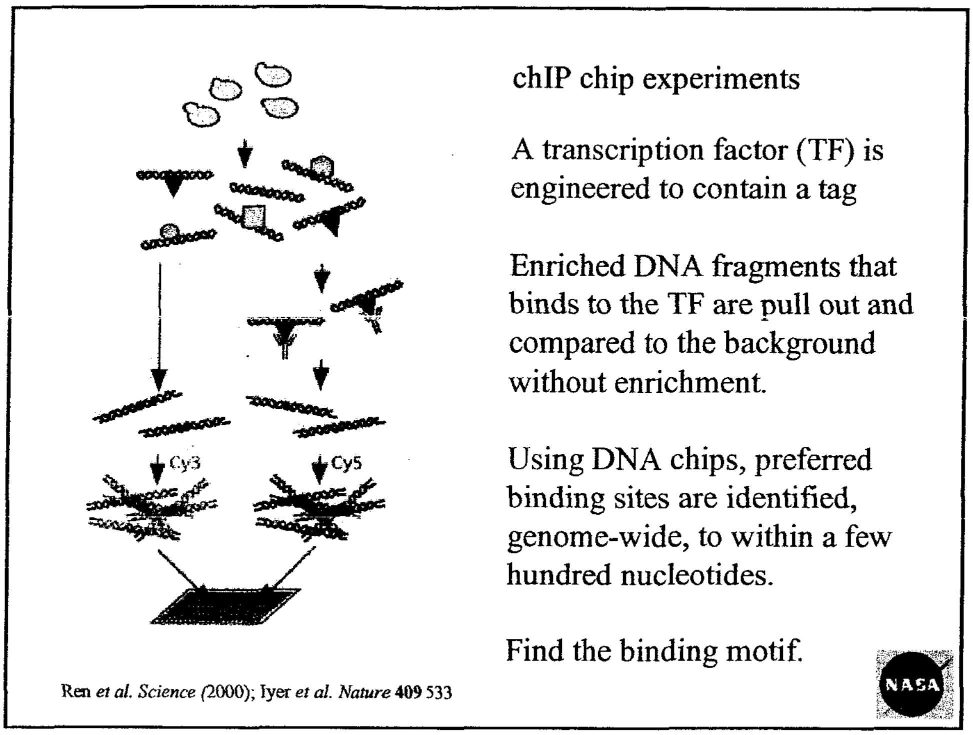
We added 50% random noise, we still recover 90% of top 2800 associations.

The method is not biased toward proteins with large interaction partners. JSN1 has the largest interaction partners, yet none of top associations involves JSN1.

summery

i) Non-random features in the genomic data are usually biologically meaningful. The key is to choose the feature well. Having a p-value based score prioritizes the findings.

ii) If two proteins share a unusually large number of common interaction partners, they tend to be involved in the same biological process. We used this finding to predict the functions of 81 un-annotated proteins in yeast.



chIP chip experiments

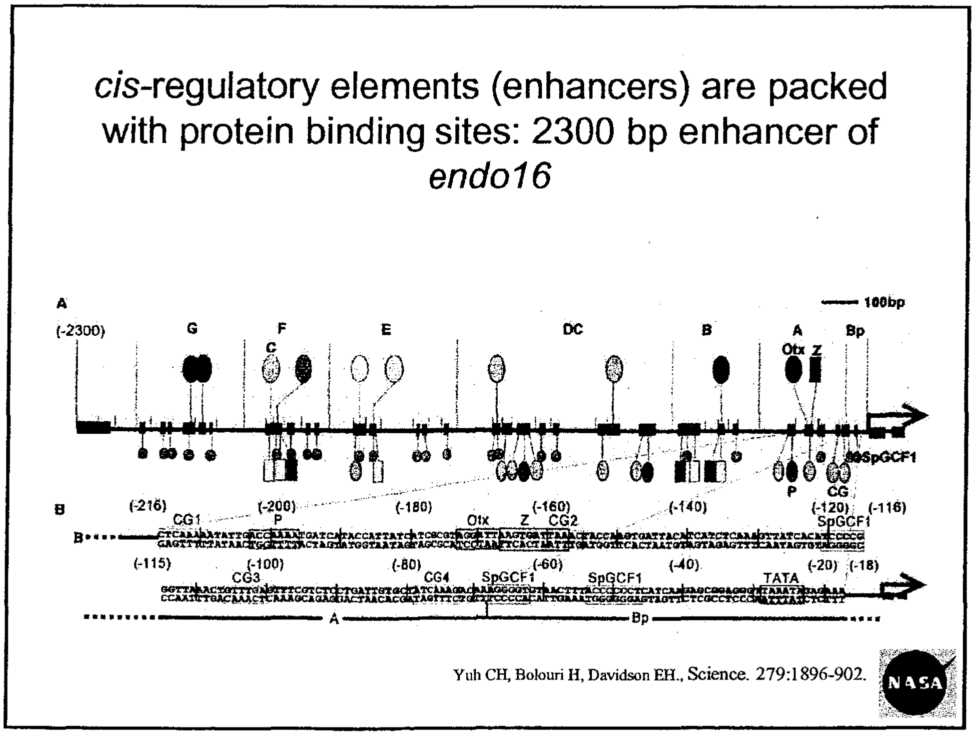
A transcription factor (TF) is engineered to contain a tag

Enriched DNA fragments that binds to the TF are pull out and compared to the background without enrichment.

Using DNA chips, preferred binding sites are identified, genome-wide, to within a few hundred nucleotides.

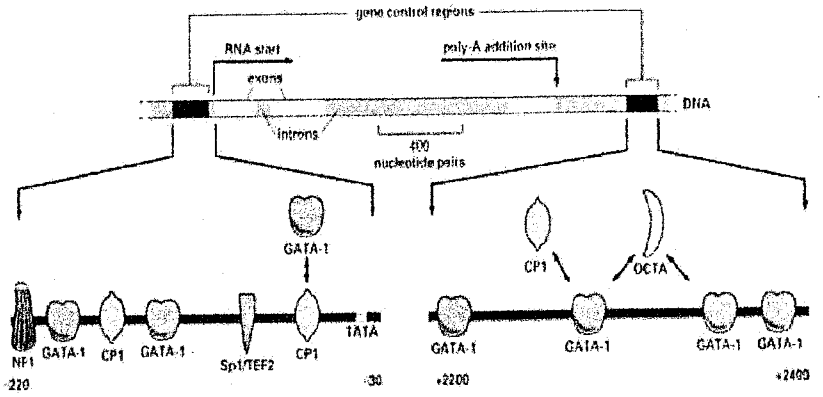
Find the binding motif.

Ren *et al. Science* (2000); Iyer *et al. Nature* **409** 533



cis-regulatory elements (enhancers) are packed with protein binding sites: 2300 bp enhancer of *endo16*

Yuh CH, Bolouri H, Davidson EH., Science. 279:1896-902.



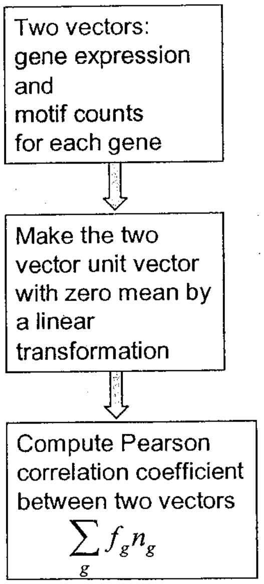
Count the number of matches to a motif pattern in the upstream region of each yeast gene

Motif: AGTT

Upstream region of gene *g*

ATCAGTTGTTGCCAGTTGTATGTCGGAGTTGTAACC

$N_g = 3$



Two vectors: gene expression and motif counts for each gene

Make the two vector unit vector with zero mean by a linear transformation

Compute Pearson correlation coefficient between two vectors $\sum_g f_g n_g$

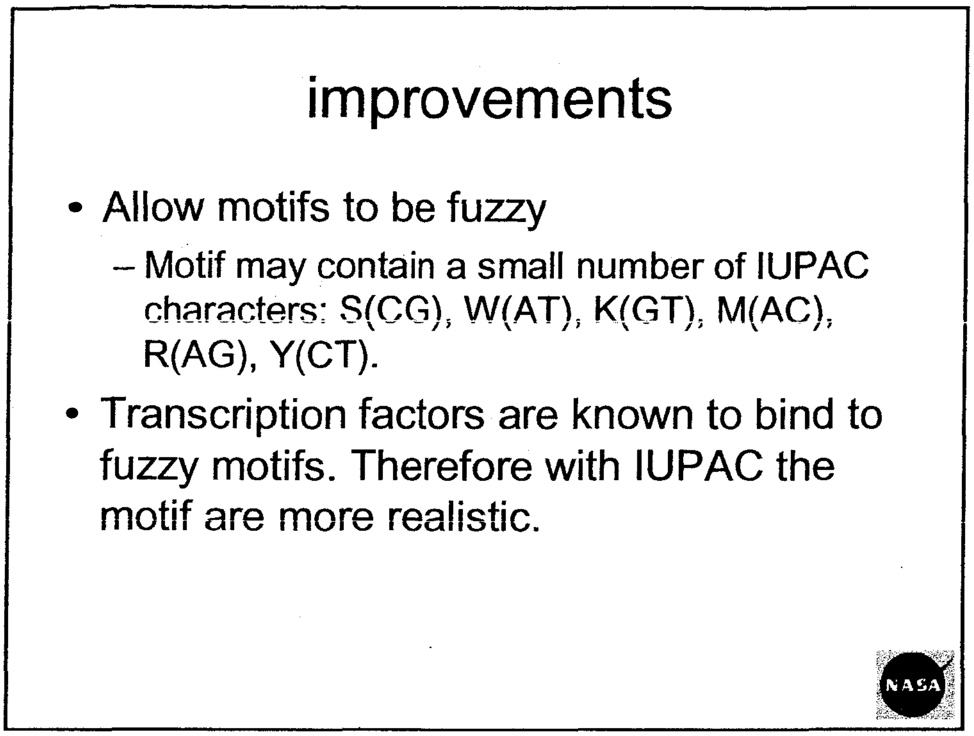
For two random unit vectors

$$\sum_g f_g n_g \approx \frac{1}{\sqrt{G}}$$

according to large number theorem

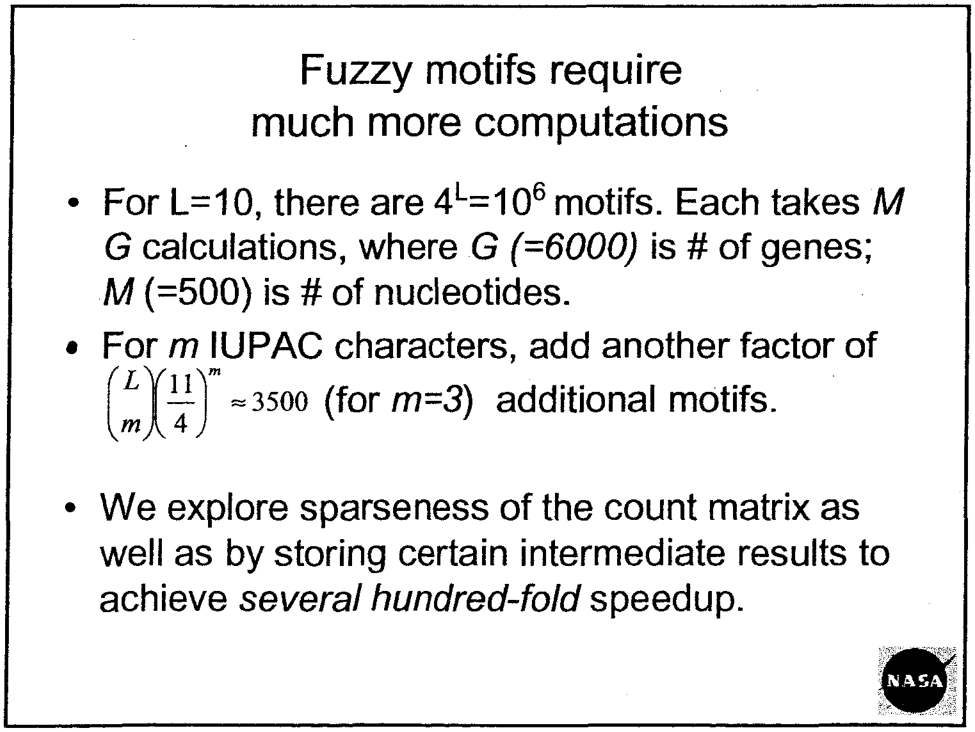
Therefore, for any motif, its correlation to gene expression can be assigned a p-value.

Bussemaker *et al. Nat. Genet.* **27 167**



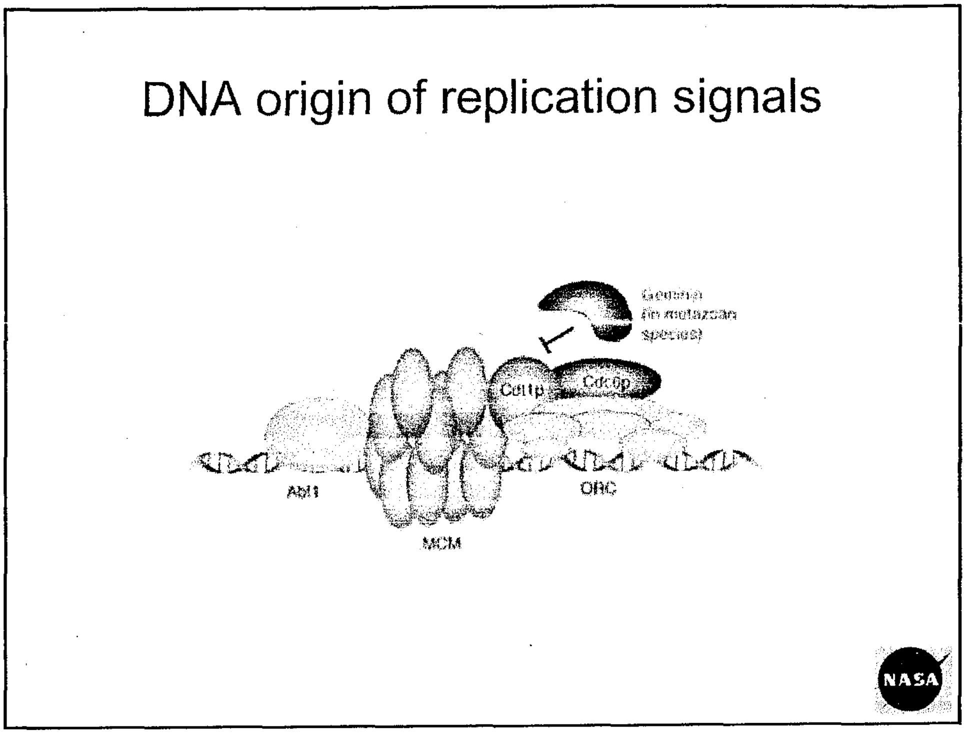
improvements

- Allow motifs to be fuzzy
  - Motif may contain a small number of IUPAC characters: S(CG), W(AT), K(GT), M(AC), R(AG), Y(CT).
- Transcription factors are known to bind to fuzzy motifs. Therefore with IUPAC the motif are more realistic.

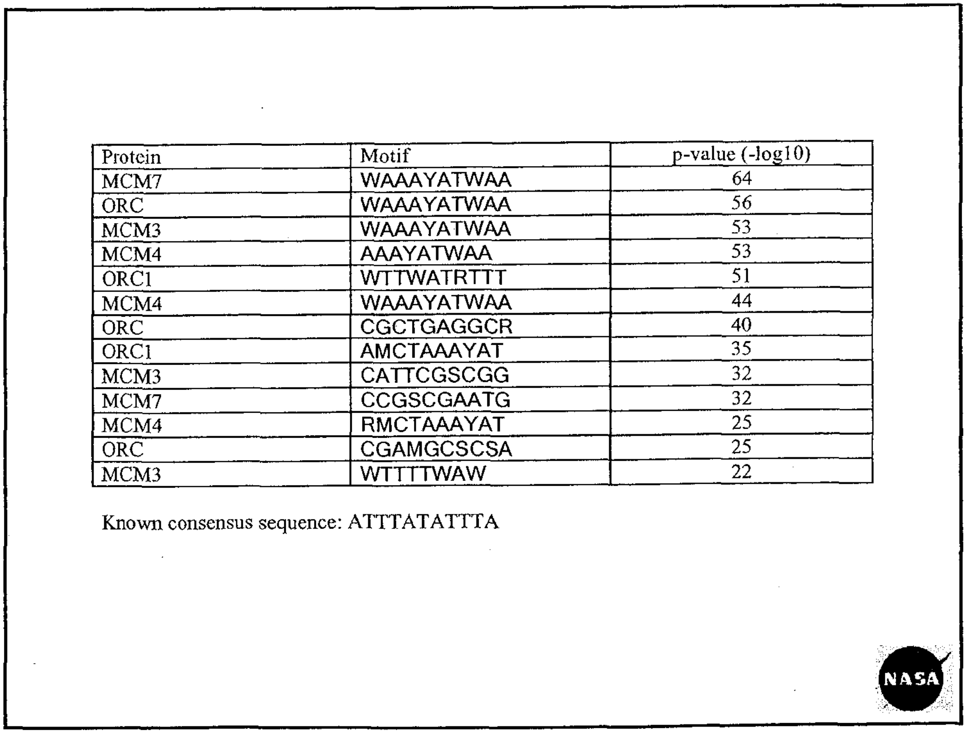


Fuzzy motifs require much more computations

- For L=10, there are $4^L=10^6$ motifs. Each takes *M G* calculations, where *G (=6000)* is # of genes; *M* (=500) is # of nucleotides.
- For *m* IUPAC characters, add another factor of $\binom{L}{m}\left(\frac{11}{4}\right)^m \approx 3500$ (for *m=3*) additional motifs.
- We explore sparseness of the count matrix as well as by storing certain intermediate results to achieve *several hundred-fold* speedup.

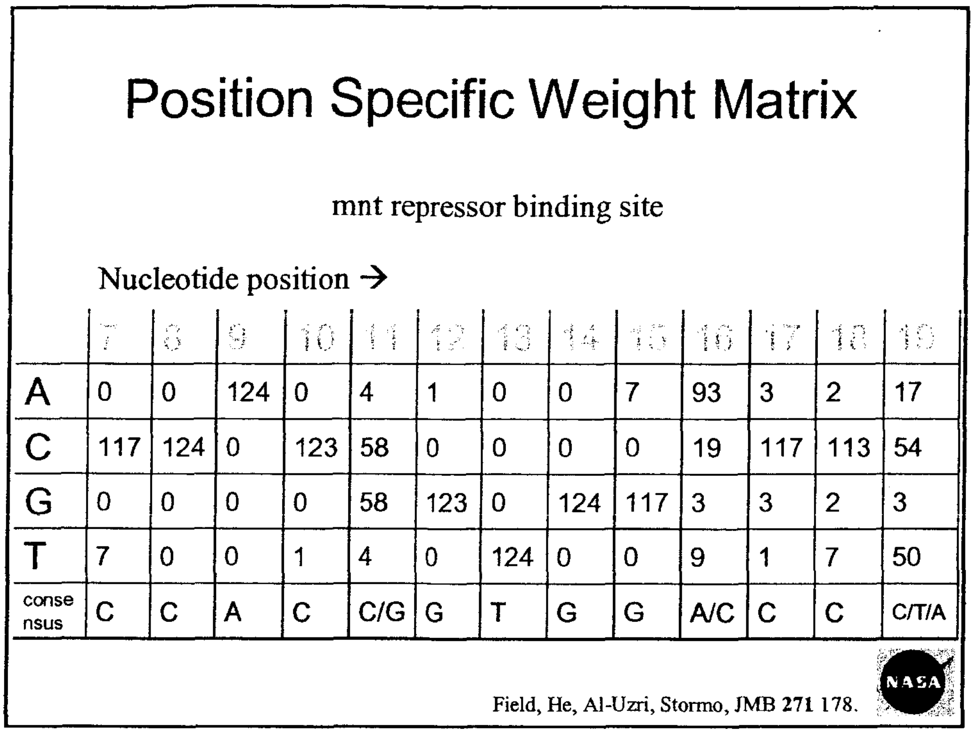


DNA origin of replication signals



Protein	Motif	p-value (-log10)
MCM7	WAAAYATWAA	64
ORC	WAAAYATWAA	56
MCM3	WAAAYATWAA	53
MCM4	AAAYATWAA	53
ORC1	WTTWATRTTT	51
MCM4	WAAAYATWAA	44
ORC	CGCTGAGGCR	40
ORC1	AMCTAAAYAT	35
MCM3	CATTCGSCGG	32
MCM7	CCGSCGAATG	32
MCM4	RMCTAAAYAT	25
ORC	CGAMGCSCSA	25
MCM3	WTTTTWAW	22

Known consensus sequence: ATTTATATTTA



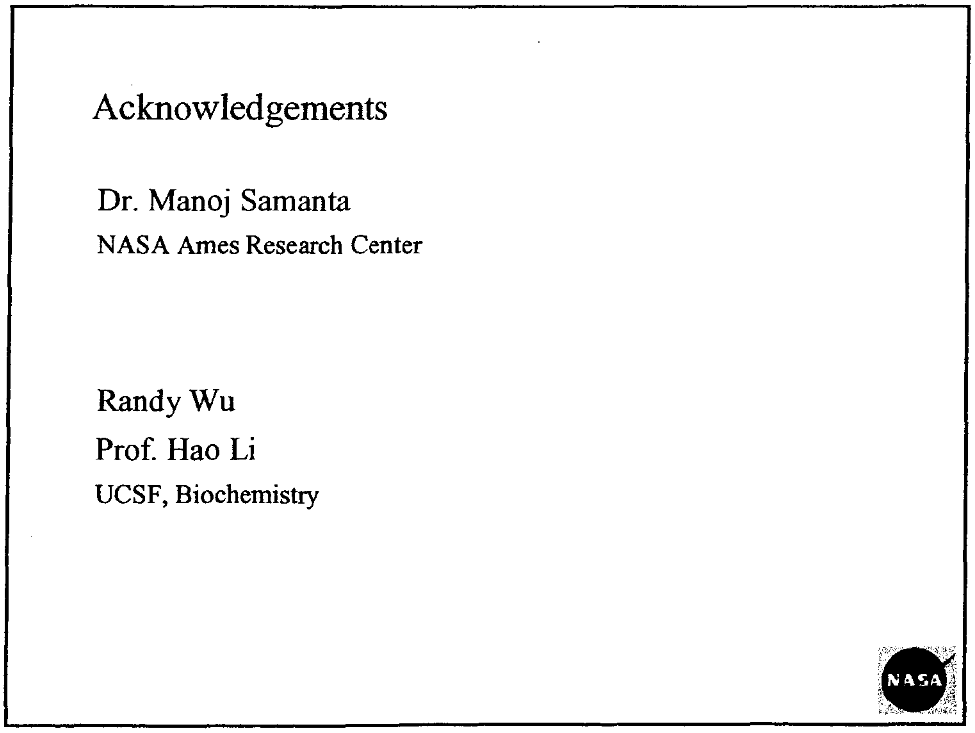
Position Specific Weight Matrix

mnt repressor binding site

Nucleotide position →

	7	8	9	10	11	12	13	14	15	16	17	18	19
A	0	0	124	0	4	1	0	0	7	93	3	2	17
C	117	124	0	123	58	0	0	0	0	19	117	113	54
G	0	0	0	0	58	123	0	124	117	3	3	2	3
T	7	0	0	1	4	0	124	0	0	9	1	7	50
consensus	C	C	A	C	C/G	G	T	G	G	A/C	C	C	C/T/A

Field, He, Al-Uzri, Stormo, JMB **271** 178.



Acknowledgements

Dr. Manoj Samanta
NASA Ames Research Center

Randy Wu
Prof. Hao Li
UCSF, Biochemistry